FACTOR ANALYSIS

- Factor analysis is a classification technique which works in an unsuperivsed learning environment.
- It is used to identify the similarity between the various features and form groups of them which it does by extracting the maximum common variance from all variables
- To perform Factor Analysis, we require continuous variable (numerical, interval scaled) with a good sample size.

Example →

- Lets have the dataset where our dependent variable is happiness which is binary categorical variable having two values, 0 and 1 denoting 0 is unhappy and 1 means happy.
- We have 20 independent variables & it is required to reduce no of features
- Suppose we get to know 20 variables comes from 4 seperate survey where each survey had 5 questions.

For example, Survey A has 5 questions. 1) How good is your salary. 2) How satisfied are you with your job 3) Do you like your workplace? 4) How understanding is your boss? 5) Do you see growth prospect in job?

Latent Variable →

- Each question is not random but are part of larger construct, known as 'Value'
- Each of these question is an 'observed variable'
- All these observed variables here represent a 'value' and this 'Value' is called 'unobserved variable' or Latent Variable and they are so called because these variable are not measured directly but rather are pointed out / indicated by observed variables.
- Thus in this example, we have 4 latent variables (4 surveys) with each latent variable having 5 observed variables.
- In real life, a statistical analysis is required to find how many values goes well together in one construct and if there is a need for having more than one construct are different from values of the other making their respective construct unique. This statiscal analysis is known as Factor Analysis.

→ Broadly we have two kind of Factor Analysis - 1) EFA (Exploratory Factor Analysis)
2) CFA (Confirmatory Factor Analysis)

- EFA is where the variables that are highly correlated to each other are grouped. Once this factor is created, it looks for another set of variables and groups them, making them another factor. The number of factors that are to be created depends and N (number of observed variables) number of factors can be created (i.e, one factor for each variable).

- CFA is used when we already have an idea about what the latent variable are and which of the observed variables belong to which latent variable. For example we have 10 variable out of which we know 5 variables are related to Education and other 5 related to sports. so here we can easily say there are two latent variables.

Extraction and Factor Rotation →

- Idea is to reduce number of variables and try to cover maximum variance (Ideal will be 100% variance). The target is reducing the maximum number of variable and some time reducing most of the variance provided by variables.
- First we need to find how many factors we need. Suppose 4. So how to group all variables into 4 factors, Factor Analysis used a method called Extraction. In this process, it find first the largest group of variables that are highly correlated to each other and creates a group from them and this group (factor) explains most of the variance of all variable in the analysis. Then it proceeds to find the next batch of highly correlated variables with this second factor explaining the second most variance in all variables and so on.

Output will look something like - Example.

Variable	Labels.	Factor 1	Factor 2	Factor 3	Factor 4
V1	How good is your salary.	0.87	0.10	0.07.	-0.06
V2	How satisfied are you with job	0.89	0.10	0.16	-0.05
V3	Do you like your workplace.	0.81	0.10	0.27	-0.06
V4	How understanding is your boss	0.~~8~~81	0.11	0.01	-0.05
V5	Do you see growth prospect in your job	0.78.	0.10	0.39	-0.07

Each of these analysis or row is factor loading. These values of these factor loading are very similar to correlation coefficient where a higher values means that variable highly defies that group (factor).

Determining Number of Factors - Use Scree Plot (Elbow method).

so factor analysis is a classification technique, we can use this technique to pick variables by forming group of correlated variables that have some meaning to them.